ISSN: 2633-4828 Vol. 5 No.4, December, 2023

International Journal of Applied Engineering & Technology

APPLICATION OF DATA MINING ALGORITHM FOR THE PREDICTION OF TUBERCULOSIS DISEASE

T. Baskar[1] and Dr. M. Kannan[2]

[1]Ph.D Research Scholar, Department of Computer Science and Applications, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram, India

[2]HOD, Department of Computer Science and Applications, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram, India

[1]basindya@gmail.com and saikannan1999@kanchiuniv.ac.in

ABSTRACT

Data mining and knowledge discovery in data is a systematic approach employed to reveal hidden patterns, associations and trends within the data. In the recent decade, data mining has been extensively utilized in several medical research investigations, encompassing the prediction and diagnosis of diseases. Datamining algorithms has the potential to offer distinct approach to aid in the diagnosis of several critical illness including tuberculosis (TB).

This paper aims to develop a model for the initial diagnosis of pulmonary tuberculosis. A preliminary diagnosis is made solely based on patient demographic information, medical history, and physical examination. Experiments were carried out utilizing classification task in machine learning. Individual classifiers like Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), C4.5 Decision Tree Classifier, K-Nearest Neighbor(KNN) Algorithm, Binary Metalogic Regression (BLR), k-mean and Apriori will be evaluated based on the parameters like accuracy, precision, sensitivity, specificity, recall and F-measure to assess the performance of the data mining model. The data were extracted from the medical records of tuberculosis patients from different hospitals in the Chennai region, Tamilnadu, India. The outcome indicated that the Apriorism technique outperformed the other single classifier in terms of all the evaluation parameters.

Keywords: Tuberculosis diagnosis; classification; Data mining techniques; Apriori

1. INTRODUCTION

Recent decades have seen an exponential growth in the generation and dissemination of bio-medical data, encompassing information gathered from pharmaceutical studies, pharmacological studies, cancer therapy investigations, research on genomes, and proteomics studies (Asha et al. 2011). Advancements in data mining techniques have resulted in the design and implementation of effective and scalable systems to extract knowledge and useful information from large datasets. Medical data mining is a prominent field of research within data mining, focusing on the analysis of vast amounts of information stored in medical databases relevant to complex clinical problems and associated diseases in patients. Prevalence of associations and similar recurring patterns within this data can uncover novel insights related with medical domain, as demonstrated in several data mining applications in medical domain.

Data classification process, utilizing information derived from past data, is a highly researched topic in statistics, decision science, and computer science. Data mining techniques have been utilized in several medical applications, such as predicting the efficacy of surgical procedures, medical testing, medication, and identifying connections among clinical and diagnostic data.

To assist clinicians in diagnosing diseases, computerized data mining and decision support tools are utilized. These tools help clinicians analyze large amounts of data from past cases to suggest a likely diagnosis based on key attributes. Many comparisons have been made between various categorization and prediction approaches, and this topic continues to be a subject of research. No single method has been universally proven to be the best for all types of data.

On a global scale, bacterial infection "tuberculosis (TB)" has resulted in a large number of fatalities when compared with infectious disease (Sánchez et al. 2009). Tuberculosis is a fatal infectious illness caused by Mycobacterium tuberculosis in humans. It often spreads via airborne transmission and affects several sections of the body, including the lungs, bones, and brain. It is a significant challenge for many poor nations due to limited access to diagnosis and treatment. Tuberculosis accounts for highest fatality rate among diseases produced by a single type of bacterium. Tuberculosis is a significant global health issue, including in India.

Various approaches, including "clinical symptoms", "tuberculin test", "sputum-smear microscopy", and "chest radiography", have been employed for diagnosing TB (Radzi et al. 2011). These methods have various limitations including being time-consuming, having low performance, difficulty in procuring sputum samples from paediatric patients, requirement of live Mycobacterium tuberculosis, needing sophisticated measurement tools operated by highly skilled medical staff, and consequently, being costly (Osman et al. 2010).

TB symptoms include "fever", "cough", "expectoration", "hemoptysis", "weight loss", and "anorexia". The symptoms are shared not just with lung cancer but also with other disorders (Bhatt et al. 2012; WHO, 2006). It results in delayed accurate diagnosis, exposure to incorrect medication, misdiagnosis, and potential death (Kusiak et al. 2000). Misdiagnosis often results from insufficient information provided by the patient or their family (Uzoka et al. 2011). A prolonged delay in diagnosing pulmonary tuberculosis hinders prompt treatment and results in the subject not being isolated. Moreover, persons who do not receive sufficient therapy are at a higher risk of developing multidrug-resistant tuberculosis (Sánchez et al. 2009).

2. LITERATURE REVIEW

Natarajan and Murthy (2011) proposed a TB classification model using "classification-based association" (CBA) and "classification based on multiple association rules" (CMAR) methodologies. The state hospital provided 300 records for the study. The dataset had twelve early symptoms and one class attribute. The suggested approach accurately identified an unknown sample as PTB or RPTB, TB with HIV. According to empirical examination, CMAR produced the best prediction rule over CBA. Asha et al. (2011) used machine learning to evaluate fundamental learning classifiers and ensembles for TB prediction. The dataset had 700 records with eleven inputs and one class attribute. After training, the models predicted pulmonary tuberculosis (PTB) and retroviral tuberculosis (RPTB), which is linked to AIDS. Using 10-fold cross-validation, classifier prediction accuracy was compared to identify the best classifier. Support Vector Machine (SVM) surpassed simple learning and random forest ensemble classifiers with 99.14% and 99.14% accuracy, respectively.

Jahantigh and Ostovare (2019) studied hidden patterns in TB patient databases. The Entropy-Shannon method identified the most important qualities, while the APRIORI method established the data association rule. R language was used to apply the proposed procedures to 548 TB patient data. The Entropy-Shannon method yielded 18 components. The APRIORI algorithm also established nine correlation criteria between maximum lift values and minimum support and confidence values. Data mining and rule extraction using APIRIORI showed that the algorithm could eliminate many rules. Nine authorized patient dataset association rules were found after eliminating these criteria. Syafrullah(2019) have applied ensemble method for the diagnosis of smear-negative pulmonary tuberculosis (SNPT). Data were extracted from the Jakarta Respiratory Center's medical records of TB patients. The findings indicated that Random Forest has recorded the highest accuracy(90.59%), this was followed by other techniques like Adaboost(90.54%) and Bagging(86.91%).

Research studies on the diagnosis of TB have utilized sound, pictures, and variables as input factors. In studies utilizing sound, researchers employed "coughing sound detection algorithms" and "lung auscultation software" that leverages lung sound waves to expedite the TB diagnosis process with high accuracy and specificity (Tracey et al. 2011; Lestari et al. 2012). Several research utilized images of TB in tissue to aid pathologists. The methods utilized include feed-forward Neural Network (Osman et al. 2009), Zernike Hybrid Moments and multilayered Perceptron Network (Osman et al. 2010), Genetic algorithm - neural network (Osman et al. 2010), compact single hidden layer feed-forward neural network (Osman et al. 2011), and hybridization signal amplification method

(Wang et al. 2011). Data mining techniques have been utilized in numerous research to identify tuberculosis based on clinical symptoms.

The present study was aimed to construct a classification model for the initial diagnosis of pulmonary TB. The input consists of patient personal data, medical history, as well as physical examination (WHO, 2006). The findings of the research can serve as a foundation for future research in the same field. The study is divided into four components.

3. RESEARCH METHOD

3.1 Classification

Classification is a data mining technique within machine learning that is utilized for predicting the group membership of data instances. Classification analysis involves categorizing data into certain classes. These methods often utilize a training set where all objects have pre-assigned class labels. The classification method utilizes the training data to generate a model. Several categorization models are used to categorize new entities, as explained below:

Linear Discriminant Analysis (LDA)

LDA algorithms are utilized in several applications including pattern recognition, and machine learning to determine a linear combination of information (Xanthopoulos et al. 2013). The concept of LDA involves calculating a linear function of the attributes to identify each class. The class function with the greatest score is considered the projected class (Balakrishnama & Ganapathiraju, 1998). It is a statistical classification algorithm that categorizes variables based on their linear combination. LDA effectively manages data when class frequencies are uneven. LDA always aims to maximize the ratio between-class variation to within-class variance in a given dataset to ensure optimal separability. Linear Discriminant study is commonly employed for classifying various biological datasets, including cancer, colon cancer, and HIV study.

SVM

Support Vector Machine (SVM) is a group of supervised learning techniques that examine data and identify patterns, primarily employed for classification tasks (Pisner & Schnyer, 2020). SVM is a non-linear classifier known for its superior classification performance when compared to other approaches. The primary concept of SVM is to create a hyperplane that serves as a decision boundary, aiming to optimize the margin between positive and negative instances. This procedure involves mapping the input sample data to a high-dimensional space in a non-linear manner, allowing for linear separation of the data and resulting in increased classification (or regression) accuracy (Pisner & Schnyer, 2020). Support Vector Machines (SVMs) are intriguing due to their strong theoretical foundation and high performance in practical applications, particularly in the field of Bioinformatics.

C4.5

C4.5 algorithm is a greedy algorithm utilized for decision tree induction. C4.5 decision trees are created using a greedy technique, constructing them in a top-down recursive divide-and-conquer method (Quinlan, 2014). Similar to ID3, constructs decision trees based on a training dataset by utilizing the principle of information entropy. The C4.5 decision tree algorithm is an advancement of the ID3 method, incorporating features such as handling missing data, continuous data, pruning, rule generation, and splitting.

K-NN

The Nearest Neighbour Classifier is a simple and direct classification algorithm in Machine Learning. It determines the class of a query example by finding its closest neighbours. This categorization approach is significant since concerns regarding bad runtime performance are less relevant due to the current computational capacity available (Cunningham & Delany, 2021). k-nearest neighbour (K-NN) algorithm is a method used to categorize objects by comparing them to the nearest training data points in the feature space [30].

Binary Logistic Regression (BLR)

The BLR is a supervised machine learning technique that models the relationship between a binary dependent variable and one or more explanatory factors using classified data after being adjusted.

K: MEAN

Clustering involves partitioning a population or dataset into distinct groups where the data points within each group are more similar to each other and dissimilar to those in other groups (Ahmed et al. 2020). K-means is essentially a classification algorithm that categorize items based on their similarities and differences.

APRIORI ALGORITHM

Association Rule Mining (ARM)

Data mining involves uncovering hidden information and intriguing patterns in databases (Borgelt, 2012). ARM is a crucial aspect of data mining that detects associations and common patterns within a group of items in specified databases. It consists of two sub-problems. 1) Identify frequent itemsets based on a predefined threshold; 2) Create association rules that meet the confidentiality constraint (Yuan, 2017).

Definitions

Here is the traditional definition of association rules. Consider a set of transactions { t1, t2,…..tn} and items I={ I1,I2,….Im }. An association rule is a statement in the form of X implies Y, where X and Y are distinct subsets of item I and have no common elements. X is referred to as the antecedent, whereas Y is known as the consequent in the rule. An itemset refers to a collection of items. Every itemset is linked to a statistical significance metric known as support. support(x) = s represents the proportion of transactions in the database that include X. The rule's level of certainty is referred to as confidence, which is calculated as the ratio of support(X Ụ Y) to support(X).

Apriori represents a significant advancement in the field of association rule mining. Apriori is efficient in the candidate generation phase. The Apriori method states that if a set of items is frequent, then all of its subsets must also be frequent (Ingle & Suryavanshi, 2015). If item set X is not large, then the item set "X" containing item sets X will never be large (Ingle & Suryavanshi, 2015). Apriori is intended to work on databases that contain transactions. The Apriori principle is advantageous since it reduces the number of items investigated by only examining item sets with a support count higher than the minimal support count.

4. Data Collection

The authors of the study personally visited different hospitals in the Chennai Region, Tamilnadu and collected data from patient records afflicted with TB. In total, 350 authentic patient records were included int the final study. All the data were consolidated into a single file containing several records. Each record contains the most pertinent information related with symptoms and specific test details of each patient. The study considered 13 symptoms (attributes) and the last attribute being designated as the class attribute (outcome variable) in associative classification. The 13 attributes included age, duration of chronic cough (weeks), symptoms of weight loss, duration of intermittent fever in days, presence of night sweats, presence of sputum, presence of blood in cough, presence of chest pain, HIV status, Diabetes Mellitus status, radiographic abnormalities, presence of wheezing, and kind of tuberculosis. Table 1 displays the names of 13 attributes and their corresponding data types.

Table 1 Dataset (N=350)

Attribute No.	Name	Datatype
1	Age	Numeric
2	Chronic cough (Weeks)	Numeric
3	Weight loss	Categorical
4	Intermittent fever (Days)	Numeric

5	Night sweats	Categorical
6	Blood cough	Categorical
7	Diabetes Mellitus (DM)	Categorical
8	Chest pain	Categorical
9	HIV	Categorical
10	Radiographic findings	Categorical
11	Sputum	Categorical
12	Wheezing	Categorical
13	TB Type	Categorical

5. Experimental Setup

The experiment utilized the open-source application Weka in various stages. Weka is a set of cutting-edge machine learning algorithms designed for various data mining tasks like data pretreatment, attribute selection, clustering, and classification. Weka has been utilized in previous studies in both clinical data mining and bioinformatics.

Weka features uses two graphical user interfaces namely Explorer and Experimenter. We can navigate between obtained results, assess models developed on various datasets, and visually represent models and datasets, including classification errors. The Experimenter feature enables automation of running classifiers and filters with various parameter settings on a dataset collection, gathering performance metrics, and doing significance tests. Experienced users can utilize the Experimenter to divide the computing burden among numerous machines utilizing Java remote method invocation.

5.1 Cross-Validation

K-fold cross-validation is a method employed in machine learning and statistical modelling to assess the effectiveness of a predictive model (Rodriguez et al. 2009). The process entails partitioning the dataset into k subsets or folds of roughly the same size. The model is trained and assessed k times, with each iteration utilizing a distinct fold as the validation set and the remaining folds as the training set (Rodriguez et al. 2009). The performance indicators from each fold are averaged to create a more reliable assessment of the model's generalization performance. In this study, the entire dataset was partitioned into 10 folds, one-fold was allocated for testing and remaining nine folds were allocated for training in the 10-fold cross-validation process.

Each instance has nine variables excluding the patient number: Age, chronic cough (weeks), weight loss, Blood cough,intermittent fever (days), Sputum, nocturnal sweats, chest discomfort, HIV, Radiographic findings, wheezing, and TB Type. The metrics are rated on an integer scale from 1 to 10, where 1 represents the least harmful and 10 the most harmful. One of the ten factors is the response variable that indicates the diagnostic status of the patient, whether they have tuberculosis (malignant) or not (benign). The training data is taken in a random manner from the complete dataset and directly fed into the suggested mining algorithm.

5.2 Evaluation Metrics for Comparative Analysis

Supervised Machine Learning (ML) utilizes many methods to assess the effectiveness of learning algorithms and the classifiers they generate. Quality metrics for categorization are derived from a confusion matrix that documents accurately and inaccurately identified instances for each category. Table 2 displays a confusion matrix for binary classification, detailing the counts of true positive (TP), false positive (FP), false negative (FN), and true negative (TN) instances.

Table 2 Confusion matrix

Actual Label	Predicted Label	
	Classified as Healthy (0)	Classified as not Healthy (1)
Actual Healthy (0)	TP	FN
Actual not Healthy (1)	FP	TN

To evaluate the performance of different classifiers various metrics were used that included sensitivity, specificity, accuracy, precision, recall and F-measure. The dataset included 308 patients with TB and 92 patients without TB disease.

Sensitivity

Sensitivity, also known as the true positive rate, is the proportion of positive instances that were correctly identified as positive.

$$Sensitivity = \frac{TP}{TP + FN} * 100\%$$

Specificity

Specificity refers to the proportion of instances that were correctly identified as negative out of all the instances that were actually negative.

$$Specificity \ \frac{TN}{FP + TN} * 100\%$$

Accuracy

Accuracy refers to the proportion of predictions that are accurate, expressed as a percentage.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} *$$

Precision

Precision is the measure of the accuracy of positive predictions, expressed as a percentage.

$$Precision = \frac{TP}{TP + FP}$$

Recall

Recall refers to the proportion of instances that were correctly identified as positive out of all the instances that were actually positive.

$$Recall \ = \frac{TP}{TP + FN}$$

F-measure

The F-measure is calculated as the harmonic mean of precision and recall.

$$F - meausre = \frac{2TP}{2TP + FP + FN} * 100\%$$

6. RESULTS AND DISCUSSION

The performance of different single classifiers on the TB dataset was evaluated using different parameters like sensitivity, specificity, accuracy, precision, recall and F-measure. Figure 1 presents the comparison of accuracy values between different classifiers (like LDA, SVM, C4.5, k-NN, BLR, k:mean and Apriori) for the TB dataset.

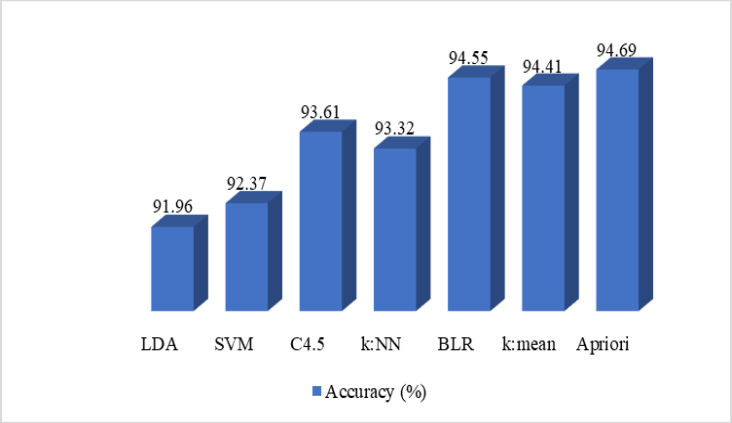

Figure 1 Accuracy of Different Classifiers on TB Dataset

The above table shows that the Apriori algorithm has produced superior performance in terms of accuracy value (94.69%) when compared with other classification approaches. Figure 2 presents the comparison of specificity values between different classifiers for the TB dataset.

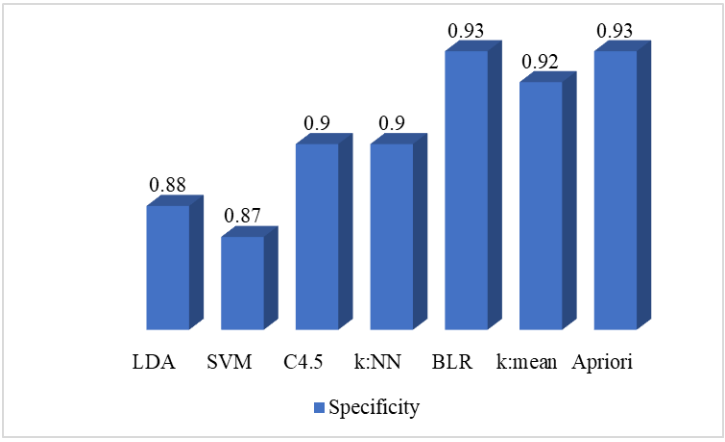

Figure 2 Specificity of Different Classifiers on TB Dataset

The above table shows that the Apriori algorithm has produced superior performance in terms of specificity value (0.93) when compared with other classification approaches. Figure 3 presents the comparison of sensitivity values between different classifiers and Apriori for the TB dataset.

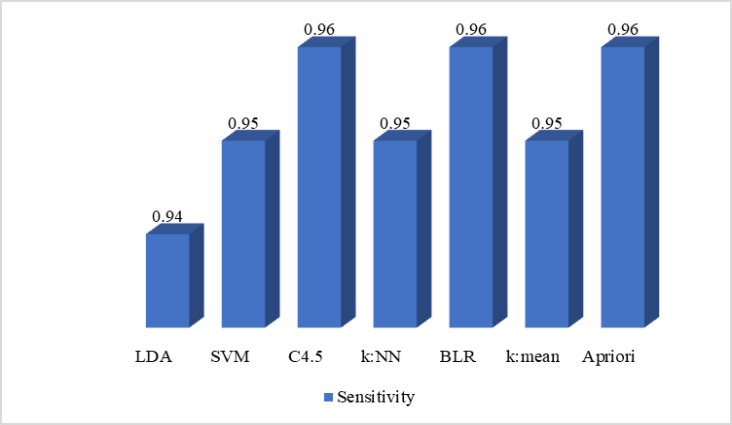

Figure 3 Sensitivity of Different Classifiers on TB Dataset

The above table shows that the Apriori algorithm has produced superior performance in terms of sensitivity value (0.96) when compared with other classification approaches. Figure 4 presents the comparison of precision values between different classifiers and Apriori for the TB dataset.

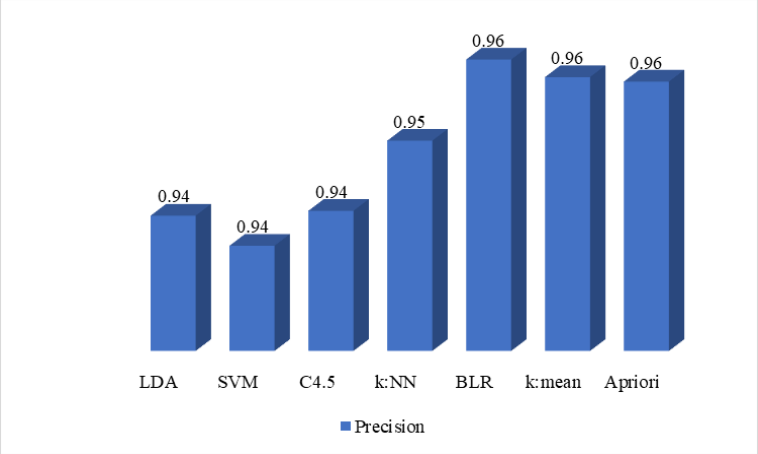

Figure 4 Precision of Different Classifiers on TB Dataset

The above table shows that the Apriori algorithm has produced superior performance in terms of precision value (0.96) when compared with other classification approaches. Figure 5 presents the comparison of recall values between different classifiers and Apriori for the TB dataset.

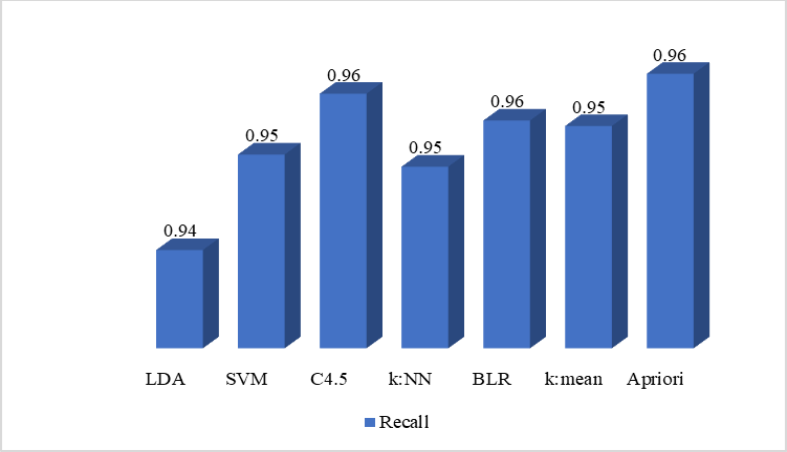

Figure 5 Recall of Different Classifiers on TB Dataset

The above table shows that the Apriori algorithm has produced superior performance in terms of recall value (0.96) when compared with other classification approaches. Figure 6 presents the comparison of F-measure values between different classifiers and Apriori for the TB dataset.

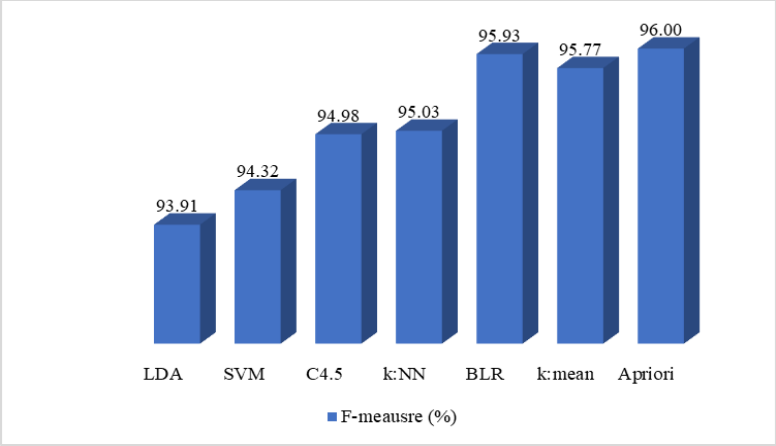

Figure 6 F-measure of Different Classifiers on TB Dataset

The above table shows that the Apriori algorithm has produced superior performance in terms of F-measure value (96.0%) when compared with other classification approaches. In addition, the computation time of Apriori algorithm was comparable and better than majority of the classifiers.

Table 3: Comparison between Different Algorithms on TB Dataset (N=350)

Algorithm	CT (ms)	Accuracy (%)	Specificity	Sensitivity	Precision	Recall	F-measure (%)
LDA	1282	91.96	0.88	0.94	0.94	0.94	93.91
SVM	1185	92.37	0.87	0.95	0.94	0.95	94.32
C4.5	1050	93.61	0.9	0.96	0.94	0.96	94.98
k:NN	985	93.32	0.9	0.95	0.95	0.95	95.03
BLR	1025	94.55	0.93	0.96	0.96	0.96	95.93
k:mean	1095	94.41	0.92	0.95	0.96	0.95	95.77
Apriori	1137	94.69	0.93	0.96	0.96	0.96	95.37

Overall, apriori algorithm has reported superior performance on the evaluation parameters like sensitivity, specificity, accuracy, precision, recall and F-measure when compared with other classifiers like LDA, SVM, C4.5, k-NN, BLR, k:mean.

7. CONCLUSION

Tuberculosis is an infectious disease that can affect all individuals. TB has been identified as the major cause of mortality in several developing nations, including India. In this research, a comparison was made on the performance of different data mining algorithms in the detection of TB using collected from different hospitals in the Chennai Region, Tamilnadu, India. The dataset consists of twelve initial symptoms (attributes) and one class attribute. The results showed that Apriori algorithm recorded best performance in terms accuracy (94.69%) and Precsion (0.96). The overall results indicate that the majority of classifier rules significantly contribute to accurately predicting tuberculosis, aiding clinicians in their diagnostic decisions. The results clearly showed that among the single classifiers considered in the study, apriori algorithm has reported the best performance in terms of majority of the evaluation parameters. Further, the performance of Apirori could be enhanced by optimization by employing ensemble methods like bagging, boosting etc.

8. Future Directions

The traditional Apriori algorithm generates significant frequent or infrequent candidate item groupings based on support count. The Apriori algorithm may need to generate a large number of candidates sets. Generating candidate sets requires multiple scans of the database. Thus, apriori requires additional memory space for the

candidate generation process. Performing repeated scans necessitates a significant amount of input/output load. To address the challenges, enhancing the Apriori algorithm through modifications is the recommended strategy. The pruning technique could be developed to reduce the number of scans needed to build candidate item sets and provide a value or weight to strong association rules. This will decrease the memory and time required to produce candidate item sets in Apriori and the algorithm will become more effective and efficient.

REFERENCES

- Agrawal, R., Srikant, R., 1994. Fast Algorithms for Mining Association Rules. In Proc. 20th int. Conf. Very Large Data Bases (VLDB), 1215, 487-499.
- Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.
- Asha, T., Natarajan, S., & Murthy, K. N. B. (2011). Effective classification algorithms to predict the accuracy of tuberculosis-A machine learning approach. *International Journal of Computer Science and Information Security*, *9*(7), 89.
- Asha, T., Natarajan, S., & Murthy, K. N. B. (2011). Effective classification algorithms to predict the accuracy of tuberculosis-A machine learning approach. *International Journal of Computer Science and Information Security*, *9*(7), 89.
- Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, *18*(1998), 1-8.
- Bhatt, M. L. B., Kant, S., & Bhaskar, R. (2012). Pulmonary tuberculosis as differential diagnosis of lung cancer. *South Asian journal of cancer*, *1*(01), 36-42.
- Borgelt, C. (2012). Frequent item set mining. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *2*(6), 437-456.
- Cunningham, P., & Delany, S. J. (2021). k-Nearest neighbour classifiers-A Tutorial. *ACM computing surveys (CSUR)*, *54*(6), 1-25.
- Ingle, M. G., & Suryavanshi, N. Y. (2015). Association rule mining using improved Apriori algorithm. *International Journal of Computer Applications*, *112*(4).
- Jahantigh, F. F., & Ostovare, M. Application of Data Mining for Exploring Hidden Patterns in Tuberculosis Patients.
- Kusiak, A., Kernstine, K. H., Kern, J. A., McLaughlin, K. A., & Tseng, T. L. (2000, May). Data mining: medical and engineering case studies. In *Industrial Engineering Research Conference* (pp. 1-7). Cleveland,
- Lestari, R., Ahmad, M., Alisjahbana, B., & Djatmiko, T. (2012, April). The lung diseases diagnosis software: Influenza and Tuberculosis case studies in the cloud computing environment. In *2012 International conference on cloud computing and social networking (ICCCSN)* (pp. 1-7). IEEE.
- Natarajan, S., & Murthy, K. N. B. (2011). A Study of Associative Classifiers with Different Rule Evaluation Measures for Tuberculosis Prediction. *IJCA Special Issue on Artificial Intelligence Techniques–Novel Approaches & Practical Applications*, 18-23.
- Osman, M. K., Ahmad, F., Saad, Z., Mashor, M. Y., & Jaafar, H. (2010, November). A genetic algorithm-neural network approach for Mycobacterium tuberculosis detection in Ziehl-Neelsen stained tissue slide images. In *2010 10th international conference on intelligent systems design and applications* (pp. 1229-1234). IEEE.

- Osman, M. K., Mashor, M. Y., & Jaafar, H. (2010, October). Detection of mycobacterium tuberculosis in Ziehl-Neelsen stained tissue images using Zernike moments and hybrid multilayered perceptron network. In *2010 IEEE international conference on systems, man and cybernetics* (pp. 4049-4055). IEEE.
- Osman, M. K., Mashor, M. Y., Jaafar, H., Raof, R. A. A., & Harun, N. H. (2009, November). Performance comparison between RGB and HSI linear stretching for tuberculosis bacilli detection in Ziehl-Neelsen tissue slide images. In *2009 IEEE International Conference on Signal and Image Processing Applications* (pp. 357-362). IEEE.
- Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning* (pp. 101-121). Academic Press.
- Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.
- Radzi, R. U. K. R. M., Mansor, W., & Johari, J. (2011, June). Review of mycobacterium tuberculosis detection. In *2011 IEEE Control and System Graduate Research Colloquium* (pp. 189-192). IEEE.
- Rodriguez, J. D., Perez, A., & Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, *32*(3), 569-575.
- Sánchez, M. A., Uremovich, S., & Acrogliano, P. (2009). Mining tuberculosis data. *Data mining and medical knowledge management: Cases and applications*, 332-349.
- Tracey, B. H., Comina, G., Larson, S., Bravard, M., López, J. W., & Gilman, R. H. (2011, August). Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis. In *2011 Annual international conference of the IEEE engineering in medicine and biology society* (pp. 6017-6020). IEEE.
- Uzoka, F. M. E., Osuji, J., Aladi, F. O., & Obot, O. U. (2011, May). A framework for cell phone based diagnosis and management of priority tropical diseases. In *2011 IST-Africa Conference Proceedings* (pp. 1-13). IEEE.
- Wang, H., Zhao, C., & Li, F. (2011, August). Identification of M. tuberculosis complex by a novel hybridization signal amplification method. In *Proceedings 2011 International Conference on Human Health and Biomedical Engineering* (pp. 1085-1088). IEEE.
- World Health Organization. (2006). Tuberculosis Coalition for Technical Assistance. *International Standards for Tuberculosis Care (ISTC). The Hague: Tuberculosis Technical Assistance*.
- World Health Organization. (2006). Tuberculosis Coalition for Technical Assistance. *International Standards for Tuberculosis Care (ISTC). The Hague: Tuberculosis Technical Assistance*.
- Xanthopoulos, P., Pardalos, P. M., Trafalis, T. B., Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear discriminant analysis. *Robust data mining*, 27-33.